ISSN 2395-8618

Robust Spoken Language Understanding for House Service Robots

Andrea Vanzo, Danilo Croce, Emanuele Bastianelli, Roberto Basili, and Daniele Nardi

***Abstract*—Service robotics has been growing significantly in the last years, leading to several research results and to a number of consumer products. One of the essential features of these robotic platforms is represented by the ability of interacting with users through natural language. Spoken commands can be processed by a Spoken Language Understanding chain, in order to obtain the desired behavior of the robot. The entry point of such a process is represented by an Automatic Speech Recognition (ASR) module, that provides a list of transcriptions for a given spoken utterance. Although several well-performing ASR engines are available off-the-shelf, they operate in a general purpose setting. Hence, they may be not well suited in the recognition of utterances given to robots in specific domains. In this work, we propose a practical yet robust strategy to re-rank lists of transcriptions. This approach improves the quality of ASR systems in situated scenarios, i.e., the transcription of robotic commands. The proposed method relies upon evidences derived by a semantic grammar with semantic actions, designed to model typical commands expressed in scenarios that are specific to human service robotics. The outcomes obtained through an experimental evaluation show that the approach is able to effectively outperform the ASR baseline, obtained by selecting the first transcription suggested by the ASR.**

***Index Terms*—Spoken language understanding, service robotics, re-ranking of automatic speech recognition systems.**

I. Introduction

DURING the recent years, the interest of the research community in the Robotics field has been rapidly increasing: robotic platforms are spreading in our domestic environments and the research on Service Robotics is becoming a hot topic. A significant aspect in this context is the study of the interaction between humans and robots, especially when this communication involves non-expert users. For this reason, natural language is a key component in human-robot interfaces. Specifically, the task of *Spoken Language Understanding* (SLU) is related to the interpretation of spoken language commands and their mapping into actions that can be executed by a robotic platform in the operational environment. Hence, the input of a typical SLU process is the user's speech, while the output can be either the corresponding action or, more in general, a response. When dealing with this problem, manifold approaches can be adopted. On the one hand, grammar-based approaches allow the design of systems that embed the entire process in a single stage, from the speech recognition up to the semantic interpretation, e.g., [1], [2], [3]. These systems rely on grammars generated by knowledge engineers, that aim at covering the (possibly vast) plethora of linguistic phenomena the user may be interested into. Moreover, these grammars can be provided with semantic attachments [4], that enable for a structured representation of the meaning of the sentence. On the other hand, approaches relying on statistical methods [5] alleviate the need to explicitly encode the information required by the NLU process, but they require training data annotated with the targeted (linguistic) phenomena the final system is expected to capture.

Regarding the Automatic Speech Recognition (ASR) systems, most of the existing off-the-shelf solutions are based on very well-performing statistical methods [6], that enable their adoption in everyday scenarios. Nevertheless, these tools rely on general-purpose language models and false positives might be generated in specific scenarios. For example, they may be optimized to transcribe queries for a Search Engine, that are characterized by different linguistic constructions with respect to a command for a robot. However, it is reasonable to expect that domain-specific scenarios provide knowledge and specific information that can improve the performance of any off-the-shelf ASR. To this regard, several works proposed techniques where a hybrid combination of free-form ASRs and grammar-based ASRs is employed to improve the overall recognition accuracy. In these approaches, the grammar-based ASR is often used to prune the transcriptions hypothesized by the free-form ASR [7], [8] or to generate new training sentences [9], [10]. Nevertheless, the above approaches are subject to several issues. In fact, as often emphasized, e.g., [11], grammar-based approaches may lack of adequate coverage, especially in dealing with the variability of (often ungrammatical) spoken language, causing a high rate of failures in the recognition of the transcription of the ASR system. On the contrary, a highly complex grammar can improve the coverage of the captured linguistic phenomena. However, this complexity may introduce ambiguities. Moreover, the cost of developing and maintaining a complex grammar may be inapplicable in realistic applications.

Manuscript received on February 8, 2016, accepted for publication on June 16, 2016, published on October 30, 2016.

Andrea Vanzo and Daniele Nardi are with the Sapienza University of Rome, Department of Computer, Control and Management Engineering "Antonio Ruberti", Rome, Italy (e-mail: {vanzo, nardi}@diag.uniroma1.it).

Danilo Croce and Roberto Basili are with the University of Roma, Tor Vergata, Department of Enterprise Engineering, Rome, Italy (e-mail: {croce, basili}@info.uniroma2.it).

Emanuele Bastianelli is with the University of Roma, Tor Vergata, Department of Civil Engineering and Computer Science Engineering, Rome, Italy (e-mail: bastianelli@ing.uniroma2.it).

ISSN 2395-8618

In this work, we propose an approach to increase the robustness of an *off-the-shelf* free-form ASR system in the context of Spoken Language Understanding for Human-Robot Interaction (HRI), relying on grammars designed over specific domains. Our target is house service robotics, with the special purpose of understanding spoken commands. We rely here on the semantic grammar proposed in [2]: this is modeled around the task of interpreting commands for robots expressed in natural language by encoding (i) the set of allowed actions that the robot can execute, (ii) the set of entities in the environment that should be considered by the robot and (iii) the set of syntactic and semantic phenomena that arise in the typical sentences of Service Robotics in domestic environment. In [2], this grammar has been used to directly provide a semantic interpretation of spoken utterances. However, this interpretation requires every sentence to be entirely recognized by this grammar: even a single word or syntactic construct missing in the process may potentially cause the failure of the overall process.

We propose here to adopt a grammar to improve the robustness of an ASR system by relying on a *scaling-down* strategy. First, we relax some of the grammar constraints allowing the coverage of shallower linguistic information. Given a grammar, we derive two lexicons designed to recognize (i) the mention to robotic actions (ii) the mention to entities in the environment. For each lexicon, we define a specific *cost* that is inversely proportional to its correctness. The transcriptions initially receive a cost that is inversely proportional to the rank provided by the ASR system and, each time one of them is recognized by the grammar or a lexicon, the corresponding cost decreases. The more promising transcription is the one minimizing the corresponding final cost. The final decision thus depends on the combination of all the costs so that, even when none of the transcriptions is recognized by the complete grammar, their rank still depends on the lexicons. In this way, those transcriptions that do not refer to any known actions and/or entities are accordingly penalized.

The proposed re-ranking strategy has been evaluated on the Human Robot Interaction Corpus (HuRIC, [12]) a collection of utterances semantically annotated and paired with the corresponding audio file. This corpus is related with the adopted semantic grammar as this has been designed by starting from a subset of utterances contained in HuRIC. Experimental results show that the proposed method is effective in re-ranking the list of hypothesis of a state-of-the-art ASR system, especially on the subset of utterances whose transcriptions are not recognized by the grammar, i.e., no pruning strategy is applicable.

In the rest of the paper, Section II provides an overview of the existing approaches to improve the quality of ASR systems. Section III presents the proposed approach and defines individual cost factors. In Section IV an experimental evaluation of the re-ranking strategy is provided and discussed. Finally, Section V derives the conclusions.

II. Related work

The robustness of Automatic Speech Recognition in domain-specific settings has been addressed in several works. In [13], the authors propose a joint model of the speech recognition process and language understanding task. Such a joint model results in a re-ranking framework that aims at modeling aspects of the two tasks at the same time. In particular, re-ranking of n-best list of speech hypotheses generated by one or more ASR engines is performed by taking the NLU interpretation of these hypotheses into account. On the contrary, the approach proposed in [14] aims at demonstrating that perceptual information can be beneficial even to improve the language understanding capabilities of robots. They formalize such information through Semantic Maps, that are supposed to synthesize the perception the robot has of the operational environment.

Regarding the combination of free-form ASR engines and grammar based systems, in [15] two different ASR systems work together sequentially: the first is grammar-based and it is constrained by the rule definitions, while the second is a free-form ASR, that is not subject to any constraint. This approach focuses on the acceptance of the results of the first recognizer. In case of rejection, the second recognizer is activated. In order to improve the accuracy of such a decision, the authors propose an algorithm that augments the grammar of the first recognizer with valid paths through the language model of the second recognizer. In [7], a robust ASR for robotic application is proposed, aiming at exploiting a combination of a Finite State Grammar (FSG) and an n-gram based ASR to reduce false positive detections. In particular, a hypothesis produced by the FSG-based decoder is accepted if it matches some hypotheses within the n-best list of the n-gram based decoder. This approach is similar to the one proposed in [16], where a *multi-pass decoder* is proposed to overcome the limitations of single ASRs. The FSG is used to produce the most likely hypothesis. Then, the n-gram decoder produces an n-best list of transcriptions. Finally, if the best hypothesis of the FSG decoder matches with at least one transcription among the n-best, then the sentence is accepted. A hybrid language model is proposed in [8]. It is defined as a combination of a n-gram model, aiming at capturing local relations between words, and a category-based stochastic context-free grammar, where words are distributed into categories, aiming at representing the long-term relations between these categories. In [9], an interpretation grammar is employed to bootstrap Statistical Language Models (SLMs) for Dialogue Systems. In particular, this approach is used to generate SLMs specific for a dialogue move. The models obtained in this way can then be used in different states of a dialogue, depending on some contextual constraints. In [17], n-grams and FSG are integrated in one decoding process for detecting sentences that can be generated by the FSG. They start from the assumption that sentences of interest are usually surrounded by carrier phrases. The n-gram is aimed at

ISSN 2395-8618

detecting those surrounding phrases and the FSG is activated in the decoding-process whenever start-words of the grammar are found.

All the above approaches can be considered complementary to the one proposed here. However, the advantages of our method are mainly in the simplicity of the proposed solution and the independence of the resulting work-flow from the adopted free-form ASR system: our aim is to define a simple yet applicable methodology that can be usable in every robot.

III. A ROBUST DOMAIN-SPECIFIC APPROACH

In this section, we propose an approach to select the most correct transcription among the results proposed by a Automatic Speech Recognition (ASR) system. Given a spoken command from the user, e.g., *move to the fridge*, such a system produces a rank of possible transcriptions such as

1) *move to the feet*
2) *more to the fridge*
3) *move to the fridge*
4) *move to the fate*
5) *move to the finch*

In this case, the correct transcription is ranked as third. In order to choose this sentence, we apply a cost function to the hypotheses based on (i) the adherence to the robot grammar, as it describes the typical commands for a robot, (ii) the recognition of action(s) applicable/known to the robot (as for *move*) and (iii) the recognition of entities, like nouns referring to objects recognized/known to the robot, e.g., *fridge*. The cost function we propose decreases along with the constraints satisfied by the sentence, e.g., the second sentence satisfies (iii), but not (i) and (ii) (as $more$ is not an action); as a consequence it results into a higher cost with respect to the third transcription. Before discussing the cost function as a ASR ranking methodology, we define the grammatical framework used in this work, in line with [2].

A. Grammar-based SLU for HRI

Robots based on speech recognition grammars usually rely on speech engines whose grammars are extended according to conceptual primitives, generally referring to known lexical theories such as Frame Semantics [18]. Early steps in the HRI chain are based on ASR modules that derive a parse tree encoding both syntactic and semantic information based on such theory. Parse trees are based on grammar rules activated during the recognition, and augmented by an instantiation of the corresponding semantic frame, that corresponds to an action the robot can execute. Compiling the suitable robot command proceeds by visiting the tree and mapping recognized frames into the final command.

The applied recognition grammar jointly models syntactic and semantic phenomena that characterize the typical sentences of HRI applications in the context of Service Robotics. It encodes a set of imperative and descriptive commands in a verb-arguments structure. Each verb is retained as it directly evokes a frame, and each (syntactic) verb argument corresponds to a semantic argument. The lexicon of arguments is semantically characterized, as argument fillers are constrained by one (or more) semantic types. For example, for the semantic argument THEME of the BRINGING frame, only the type TRANSPORTABLE_OBJECTS is allowed. As a consequence, a subset of words referring to things transportable by the robots, e.g., *can*, *mobile phone*, *bottle* is accepted. A subset of the grammar for the BRINGING frame, covering the sentence *Bring the book to the table* is reported hereafter:

`Bringing` → `Target Theme Goal` | ...
`Target` → *bring* | *carry* | ...
`Theme` → *the* `Transportable_objects` | ...
`Transportable_objects` → *can* | *book* | *bottle* | ...
`Goal` → ...

We will distinguish between terminals denoting entities (such as can, $book$, $bottle$ that belong to the lexicon of TRANSPORTABLE_OBJECTS) from the lexicon of possible actions (such as $bring$, $take$ or $carry$ characterizing the actions of the frame BRINGING) as they will give rise to different predicates augmented with grammatical constraints. Moreover, transcribed sentences covered by the grammar, i.e., belonging to the grammar language, are more likely to correspond to the intended command expressed by the user, and should be ranked first in the ASR output.

B. A grammar-based cost model for accurate ASR ranking

A first interesting type of constraint is posed by the ASR system itself. In fact the rank proposed by an ASR system is usually driven by a variety of linguistic knowledge in the ASR device. A basic notion of cost can be thus formulated ignoring the domain of the specific grammar.

Given a spoken utterance v, let $\mathcal{H}(v)$ be the corresponding list of hypotheses produced by the ASR. The size $|\mathcal{H}(v)| = N$ corresponds to the number of hypotheses. Each hypothesis $h \in \mathcal{H}(v)$ is a pair $\langle s, \omega(s) \rangle$, where s is the transcription of v, and $\omega(s)$ is a cost attached to s. Let $p(s)$ be its position in the ASR systems ranking. According to this cost function, the higher is $\omega(s)$, the lower the confidence in h being the correct transcription.

Since many off-the-shelf ASR systems do not provide the confidence score for each transcription, in order to provide a general solution, only the rank is taken into account: let v be a spoken utterance and $\mathcal{H}(v)$ the corresponding list of transcriptions, then, $\forall s \in \mathcal{H}(v)$ the *ranking cost* ω_{rc} is defined as follows:

$$\omega_{rc}(s, \theta) = \frac{p(s) + \theta}{\sum_{s' \in \mathcal{H}(v)} p(s') + \theta N} \tag{1}$$

where $p(s)$ corresponds to the position $(1, \ldots, |H(v)|)$ of s in $\mathcal{H}(v)$. Here θ is a smoothing parameter that enables the tuning of the variability allowed to the final rank with respect to the initial rank proposed by the ASR system.

ISSN 2395-8618

The overall cost assigned to a transcription s depends on the ASR ranking as well as on the grammar. Let $s \in \mathcal{H}(v)$, let ω_i be a parametric cost depending on the grammar $\mathcal{G}$, the overall cost $\omega(s)$ can be defined as:

$$\omega(s) = log\big(\omega_{rc}(s,\theta)\big) + \sum_i log\big(\omega_i(s,\alpha_i)\big) \tag{2}$$

where the different ω_i capture different aspects of the grammar $\mathcal{G}$ with scores derived from the grammatical or lexical criteria. Higher values of ω_i correspond to stronger violations. Moreover, $\omega_{rc}(s,\theta)$ is the ranking cost as in Equation (1), while α_i is the parameter associated to each cost ω_i.

In this paper we investigate three possible cost factors, i.e., $i = 1, 2, 3$, to enforce information derived by different grammatical, i.e., domain-dependent, constraints. As these can be different, we designed three different cost factors:

- $\omega_G(s, \alpha_G)$ is the *complete-grammar cost* that is minimal when the transcription belongs to the language generated by the grammar $\mathcal{G}$, and maximal otherwise;
- $\omega_A(s, \alpha_A)$ is the *action-dependent cost* that is minimal when the transcription explicitly refers to actions the robot is able to perform, and maximal otherwise;
- $\omega_E(s, \alpha_E)$ is the *entity-dependent cost* that takes into account the entities targeted by the commands, and is minimal if they are referred into the transcription s and maximal otherwise.

These cost factors are detailed hereafter.

Complete-grammar cost. When dealing with the Spoken Language Understanding with robots we may want to restrict the user sentences to a set of possible commands. This is often realized by defining a grammar covering the linguistic phenomena we want to catch. Moreover, if the grammar is designed to embed also semantic information as in [2], it can be introduce also higher level semantic constraints. For instance, the BRINGING action can be applied only to TRANSPORTABLE_OBJECTS. As an example, a sentence a transcription such as *bring me the fridge* is discarded by the grammar if the *fridge* is not a TRANSPORTABLE_OBJECTS.

Let $\mathcal{G}$ be a grammar designed for parsing commands for a robot R. Let $L(\mathcal{G})$ be the language generated by the grammar, i.e., the set of all possible sentences that $\mathcal{G}$ can produce. Then, the *complete-grammar cost* ω_G is computed as

$$\omega_G(s, \alpha_G) = \begin{cases} \alpha_G & \text{if } s \in L(\mathcal{G}) \\ 1 & \text{otherwise} \end{cases} \tag{3}$$

where $\alpha_G \in (0, 1]$ is a weight that measures the strength of the violation and can be used to weight the impact of an *"out-of-grammar"* transcription. Notice that the weight α_G can be either set as a subjective confidence or tuned through a set of manually validated hypotheses. If α_G is set to 1, no grammatical constraint is applied and the complete grammar cost has no effect.

Action-dependent cost. Robot specifications enable the construction of the lexicon of potential actions A, hereafter called $\mathcal{L}_A$. Let A be the set of actions that a robot can perform, e.g., MOVE, GRASP, OPEN. For each action $a \in A$, a corresponding set of lexical entries can be used to linguistically refer to a: we will denote such a set as $\mathcal{L}(a) \subset \mathcal{L}_A$.

The *action-dependent cost* ω_A for a transcription $s \in \mathcal{H}(v)$ is thus given by:

$$\omega_A(s, \alpha_A) = \prod_{\forall w \in s} \alpha_A(w) \tag{4}$$

where $\alpha_A(w)$ is defined as:

$$\alpha_A(w) = \begin{cases} \alpha_A & \exists a \in A \text{ such that } w \in \mathcal{L}(a) \\ 1 & \text{otherwise} \end{cases} \tag{5}$$

$\alpha_A \in (0, 1]$ is a weight that favors words corresponding to actions that are in the repertoire of the robot. The weight α_A can be either set as a subjective preference or tuned over a set of manually validated hypotheses. Note that if α_A is set to 1, no action dependent constraint is applied and the corresponding cost is not triggered.

Entity-dependent cost. Exploiting environment observations can be beneficial in interpreting commands. Notice that the objects of the robot's environment are more likely to be referred by correct transcriptions rather than by the wrong ones, as these are usually "out of scope". Let $\mathcal{G}$ be the grammar designed for commands. Given the set of terminals of $\mathcal{G}$, in the lexicon $\mathcal{L}_G$ a specific set of terms is used to make (explicit) reference to objects of the environment. For each entity e (e.g., MOVABLE OBJECTS such as $bottles$, $books$, ..., or FURNITURES, such as $table$ or $armchair$) the set of nouns used to refer to e in the language $L(\mathcal{G})$ is well defined, and it is denoted by $\mathcal{L}(e)$.

The *entity-dependent cost* ω_E for a transcription $s \in \mathcal{H}(v)$ is thus given by:

$$\omega_E(s, \alpha_E) = \prod_{\forall w \in s} \alpha_E(w) \tag{6}$$

where $\alpha_E(w)$ is defined as:

$$\alpha_E(w) = \begin{cases} \alpha_E & \exists \text{ entity } e \text{ such that } w \in \mathcal{L}(e) \\ 1 & \text{otherwise} \end{cases} \tag{7}$$

and $\alpha_E \in (0, 1]$ is a weight that favors words corresponding to entities the robot is able to recognize in the environment. The weight α_E can be either set as a subjective preference or tuned over a set of manually validated hypotheses. Also α_E, when set to 1, produces no entity dependent constraint and corresponds to a null impact on the final cost.

IV. EXPERIMENTAL EVALUATIONS

The grammar employed in these evaluations has been designed in [19], lately improved in [2], and its definition is compliant to the Speech Recognition Grammar Specification [4]. The grammar takes into account 17 frames, each of which is evoked by an average of 2.6 lexical units.

ISSN 2395-8618

On average, for each frame 27.9 syntactic patterns are defined. Entities are clustered in 28 categories, with an average amount of items per cluster of 11.2 elements. We extracted an Actions Lexicon $\mathcal{L}(a)$ containing 44 different verbs. The Entities Lexicon $\mathcal{L}(e)$ is composed of 216 and 97 single and compound words, respectively, with a total amount of 313 entities. The dataset of the empirical evaluation is the HuRIC corpus[1], a collection of utterances annotated with semantic predicates and paired with the corresponding audio file. HuRIC is composed of three different datasets, that display an increasing level of complexity in relationship with the grammar employed.

The **Grammar Generated** dataset (GG) contains sentences that have been generated by the above speech recognition grammar. The **Speaky for Robot** dataset (S4R) has been collected during the *Speaky for Robots project*[2] and contains sentences for which the grammar has been designed, so that the grammar is supposed to recognize a significant number of utterances. While the grammar is expected to cover all the sentences in the GG dataset, this may be not true for the S4R one, as some sentences are characterized by linguistic structures not considered in the grammar definition. The **Robocup** dataset (RC) has been collected during the 2013 *Robocup@Home* competition [20] and it represents the most challenging section of the corpus, given its linguistic variability. In fact, even referring to the same house service robotics, it contains sentences not constrained by the grammar structure, as, during the acquisition process, speakers were allowed to say any kind of sentence related to the domain.

The experimental evaluation aimed at measuring the effectiveness of the approach we proposed. To this end, the cost function $\omega(s)$ has been used in different settings. The α_i can be used to properly activate/deactivate the costs operating on specific evidences. In fact, if $\alpha_i = 1$, the corresponding cost is not triggered. However, whenever a cost is activated, its parameter has been estimated through 5-fold cross validation (with one fold for testing), as well as the θ smoothing parameter of the ranking cost ω_{rc}. Performances have been measured in terms of Precision at 1 (P@1), that is the percentage of correctly transcribed sentences occupying the first position in the rank, and Word Error Rate (or WER). All audio files are analyzed through the official Google ASR APIs [21]. In order to reduce the evaluation bias to ASR errors, only those commands with an available solution within the 5 input candidates were retained for the experiments.

A. Experimental Results

Table I shows the mean and standard deviation of the P@1 and the WER across the 5 folds. The results have been obtained by testing our cost function on the aforementioned HuRIC corpus. The transcription have been gathered in January 2016. The sizes of the GG, S4R and RC datasets were of 100, 97 and 112 utterances, each paired with 5 transcriptions derived from the ASR system.

TABLE I
RESULTS IN TERMS OF P@1 AND WER

	GG		S4R		RC	
	P@1	WER	P@1	WER	P@1	WER
ASR BL	74.00 ±6.52	3.66	84.71 ±7.57	2.61	79.55 ±10.66	3.89
Greedy	**94.79** ±0.12	4.33	93.58 ±4.43	1.09	79.30 ±7.96	5.00
ω_G	90.00 ±3.54	1.13	**93.98** ±6.36	0.89	78.64 ±9.59	3.92
ω_A	80.00 ±7.07	2.22	82.71 ±10.02	2.85	82.27 ±10.21	3.65
ω_E	78.00 ±5.70	2.97	83.66 ±6.04	3.00	**83.18** ±11.32	3.19
$\omega_{G,A}$	90.00 ±3.54	1.13	92.93 ±6.63	1.06	80.45 ±11.54	3.79
$\omega_{E,G}$	90.00 ±3.54	1.13	**93.98** ±6.36	0.89	82.27 ±10.71	3.23
$\omega_{A,E}$	83.00 ±2.74	1.94	86.72 ±5.42	2.21	**83.18** ±10.85	3.71
$\omega_{G,A,E}$	90.00 ±3.54	1.13	92.93 ±6.63	1.06	82.27 ±12.07	3.75

We compared our approach, where hypotheses are re-ranked according to our cost function $\omega(s)$, to two different baselines. In the first baseline (*ASR BL*), the best hypothesis is selected by following the initial guess given by the ASR, i.e., the transcription ranked in first position. The second baseline (*Greedy*) selects the first transcription, occurring within the list, that belongs to the language generated by the grammar. Conversely, the row ω_G refers to the cost function setting when α_A and α_E are set to 1, i.e., just the cost ω_G is actually triggered. In general, $\omega_{i,j,k}$ refers to the cost function when the costs ω_i, ω_j and ω_k are considered.

The *Greedy* approach seems to be effective when the sentences are more constrained by the grammar, i.e., it is likely that the correct transcription is recognized by the grammar. In fact, this approach is able to reach high scores of P@1 in both GG and S4R datasets, i.e., 94.79 and 93.58, respectively. Moreover, when the *complete-grammar cost* is triggered, i.e., ω_G, $\omega_{G,A}$ and $\omega_{G,A,E}$, we get comparable results, specially on the S4R dataset, with a relative increment of $+10.94\%$. These observations do not apply for the RC dataset, where the structures and lexicon of the sentences are not constrained by the grammar. In fact, the *complete-grammar cost* does not seem to provide any actual improvement.

Conversely, we observe a drop of performance when the full constrained grammar is employed, i.e., both *Greedy* and ω_G. On the other hand, when the *action-dependent* and *entity-dependent costs* are considered, we reach the best results. In particular, ω_E and $\omega_{A,E}$ are able to outperform both the *ASR BL* and the grammar constrained approaches. This behavior seems to depict a sort of *scaling-down* strategy: when the grammar does not fully cover the sentence, or it is not available, we can still rely on simpler, but more effective, information. Nevertheless, even though it does not perform the best, the strategy where all costs are triggered, i.e., $\omega_{G,A,E}$, seems to be the most stable across different sentence complexity conditions.

We conducted experiments on the transcription lists that have been employed in [14]. These have been gathered by relying on the same ASR engine, but almost two years earlier (May 2014). Hence, a different amount of sentences are employed in this experiment. In fact, the GG, S4R and RC

[1] http://sag.art.uniroma2.it/demo-software/huric/
[2] http://www.dis.uniroma1.it/~labrococo/?q=node/3

ISSN 2395-8618

TABLE II
RESULTS IN TERMS OF $P@1$ AND WER OBTAINED OVER DATA USED IN [14]

	GG		S4R		RC	
	P@1	WER	P@1	WER	P@1	WER
ASR BL	84.18 ±11.53	2.04	85.48 ±6.80	4.61	78.75 ±8.39	5.15
Greedy	**94.00** ±5.48	2.36	**95.78** ±5.79	0.62	74.96 ±5.33	5.80
ω_G	92.00 ±8.37	0.74	92.60 ±5.48	2.09	80.00 ±8.15	4.82
ω_A	86.00 ±13.42	1.47	85.48 ±6.80	4.30	82.50 ±6.85	2.69
ω_E	84.18 ±11.53	2.04	82.40 ±6.41	3.37	83.75 ±3.42	3.57
$\omega_{G,A}$	92.00 ±8.37	0.74	92.88 ±7.05	1.41	82.50 ±5.23	2.66
$\omega_{G,E}$	92.00 ±8.37	0.74	92.60 ±5.48	2.09	82.50 ±5.23	2.98
$\omega_{A,E}$	86.00 ±13.42	1.47	83.94 ±7.84	3.32	**90.00** ±8.39	1.85
$\omega_{G,A,E}$	92.00 ±8.37	0.74	92.88 ±7.05	1.41	83.75 ±3.42	2.66

datasets are composed of 51, 68 and 80 lists, respectively. The results are shown in Table II. We observe here similar trend, with both *Greedy* and *complete-grammar cost* reaching the highest scores in GG and S4R datasets. Even though the results obtained on these corpora are still comparable with the ones presented in [14], the interesting behavior observed on the RC dataset represents the main substantial difference. Even on this dataset, the trend seems to be the same, with the $\omega_{A,E}$ outperforming any other approach with relative improvements in P@1 up to $+20.06\%$. The trend of $\omega_{G,A,E}$ is confirmed here, making it the best solution as the most stable approach.

V. CONCLUSIONS

In this work, we presented a practical approach to increase the robustness of an *off-the-shelf* free-form Automatic Speech Recognition (ASR) system in the context of Spoken Language Understanding for Human-Robot Interaction (HRI), relying on grammars designed over specific domains. In particular, a cost is assigned to each ASR transcription, that decreases along with the number of constraints satisfied by the sentence with respect to adopted grammar. Despite to the simplicity of the proposed method, experimental results show that the proposed method allows to significantly improve a state-of-the-art ASR system over a dataset of spoken commands for robots.

Future work will consider the adoption of this re-ranking strategy within full chains of Spoken Language Understanding in the context of HRI, as the one presented in [5]. Moreover, the simple proposed method can be jointly used with supervised learning methods ([14]) that may exploit evidenced derived from the grammar to learn more expressive re-ranking functions.

REFERENCES

[1] J. Dowding, J. M. Gawron, D. Appelt, J. Bear, L. Cherny, R. Moore, and D. Moran, "Gemini: A natural language system for spoken-language understanding," AI Center, SRI International, 333 Ravenswood Ave., Menlo Park, CA 94025, Tech. Rep. 527, Apr 1993, presented at the 31st Annual Meeting of the Association for Computational Linguistics, 22-26 June 1993, Columbus, OH.

[2] E. Bastianelli, D. Nardi, L. C. Aiello, F. Giacomelli, and N. Manes, "Speaky for robots: the development of vocal interfaces for robotic applications," *Applied Intelligence*, vol. 44, no. 1, pp. 43–66, 2015.

[3] G.-J. Kruijff, H. Zender, P. Jensfelt, and H. I. Christensen, "Situated dialogue and spatial organization: What, where... and why?" *International Journal of Advanced Robotic Systems*, vol. 4, no. 1, pp. 125–138, Mar 2007, special Issue on Human and Robot Interactive Communication.

[4] A. Hunt and S. McGlashan, "Speech recognition grammar specification," World Wide Web Consortium, Tech. Rep., 2004.

[5] E. Bastianelli, G. Castellucci, D. Croce, R. Basili, and D. Nardi, "Effective and robust natural language understanding for human-robot interaction," in *Proceedings of 21st European Conference on Artificial Intelligence*. IOS Press, 2014, pp. 57–62.

[6] G. Hinton, L. Deng, D. Yu, A. rahman Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. S. G. Dahl, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, November 2012.

[7] M. Doostdar, S. Schiffer, and G. Lakemeyer, *RoboCup 2008: Robot Soccer World Cup XII*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, ch. A Robust Speech Recognition System for Service-Robotics Applications, pp. 1–12.

[8] D. Linares, J.-M. Benedí, and J.-A. Sánchez, "A hybrid language model based on a combination of n-grams and stochastic context-free grammars," *ACM Transactions on Asian Language Information Processing*, vol. 3, no. 2, pp. 113–127, Jun 2004.

[9] R. Jonson, "Grammar-based context-specific statistical language modelling," in *Proceedings of the Workshop on Grammar-Based Approaches to Spoken Language Processing*, ser. SLP 2007, Stroudsburg, PA, USA, 2007, pp. 25–32.

[10] H. Li, T. Zhang, R. Qiu, and L. Ma, "Grammar-based semi-supervised incremental learning in automatic speech recognition and labeling," *Energy Procedia*, vol. 17, Part B, pp. 1843–1849, 2012, 2012 International Conference on Future Electrical Power and Energy System.

[11] R. de Mori, "Spoken language understanding: a survey," in *IEEE Workshop on Automatic Speech Recognition & Understanding, ASRU 2007, Kyoto, Japan, December 9-13, 2007*, S. Furui and T. Kawahara, Eds. IEEE, 2007, pp. 365–376.

[12] E. Bastianelli, G. Castellucci, D. Croce, L. Iocchi, R. Basili, and D. Nardi, "HuRIC: A human robot interaction corpus," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*. Reykjavik, Iceland: European Language Resources Association (ELRA), May 2014.

[13] F. Morbini, K. Audhkhasi, R. Artstein, M. Van Segbroeck, K. Sagae, P. Georgiou, D. Traum, and S. Narayanan, "A reranking approach for recognition and classification of speech input in conversational dialogue systems," in *Spoken Language Technology Workshop (SLT), 2012 IEEE*, Dec 2012, pp. 49–54.

[14] E. Bastianelli, D. Croce, R. Basili, and D. Nardi, "Using semantic maps for robust natural language interaction with robots," in *Sixteenth Annual Conference of the International Speech Communication Association*. International Speech Communication Association, 2015, pp. 1393–1397.

[15] M. Levit, S. Chang, and B. Buntschuh, "Garbage modeling with decoys for a sequential recognition scenario," in *IEEE Workshop on Automatic Speech Recognition Understanding, ASRU 2009*, 2009, pp. 468–473.

[16] S. Heinrich and S. Wermter, "Towards robust speech recognition for human-robot interaction," in *Proceedings of the IROS2011 Workshop on Cognitive Neuroscience Robotics (CNR)*, Sep 2011, pp. 23–28.

[17] Q. Lin, D. Lubensky, M. Picheny, and P. S. Rao, "Key-phrase spotting using an integrated language model of n-grams and finite-state grammar," in *EUROSPEECH*, G. Kokkinakis, N. Fakotakis, and E. Dermatas, Eds. ISCA, 1997.

[18] C. J. Fillmore, "Frame semantics and the nature of language," *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*, vol. 280, no. 1, pp. 20–32, 1976.

[19] L. C. Aiello, E. Bastianelli, L. Iocchi, D. Nardi, V. Perera, and G. Randelli, "Knowledgeable talking robots," in *Artificial General Intelligence - 6th International Conference, AGI 2013, Beijing, China, July 31 - August 3, 2013 Proceedings*, 2013, pp. 182–191.

[20] T. Wisspeintner, T. van der Zant, L. Iocchi, and S. Schiffer, "RoboCup@Home: Scientific competition and benchmarking for domestic service robots," *Interaction Studies*, vol. 10, no. 3, pp. 392–426, 2009.

[21] C. Chelba, P. Xu, F. Pereira, and T. Richardson, "Distributed acoustic modeling with back-off n-grams," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar 2012, pp. 4129–4132.